R&T PCIe400 : Generic DAQ

Julien Langouët *(CPPM) on behalf of the R&T PCIe400 team*
CPPM, IJClab, LP2i, LAPP, LPCC, LHCb Online

Outline

Context

Technical development

High bandwidth bus

Organization

Conclusion

Context

DAQ architecture LHCb upgrade II

Requirements comparison to Upgrade I

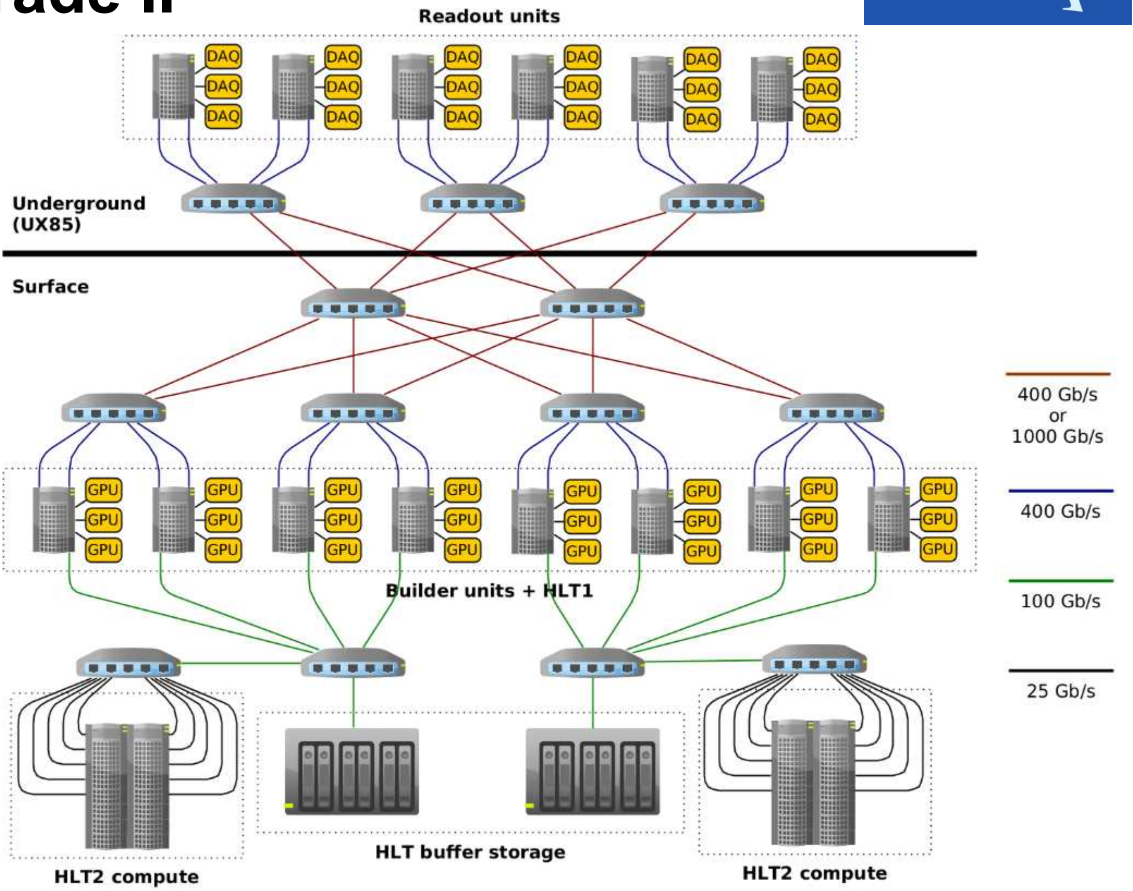
- **3x** number of data links
  11 000 → 33 000
- **2x** bandwidth of front-end serial links
  5 Gbit/s → 10 Gbit/s
- **5x** data throughput
  32 Tbit/s → 160 Tbit/s

Similar architecture to Upgrade I

- Generic readout DAQ card
- Aggregate high number of front end link → 1 commercial protocol link

New « split » approach

- Readout units separated from event builder



Envisaged DAQ architecture LHCb upgrade II

PCIe400 rationale

Detector upgrade

- New ECAL, RICH and Downstream Tracker* will be deployed in LS3 (2026-2028)
  - Adoption of the common new serializer Versatile Link+ (protocol lpGBT)
  - 4D techniques require low jitter and phase determinism $\mathcal{O}$(10)ps RMS on the master clock

Experimental path for new DAQ system architecture

- Integrate a network interface (400GbE) to reduce cost and increase flexibility
- Implement cache memory coherent protocol to use board as an acceleration card

PCIe400 is therefore a fundamental development to keep pace with technology evolution

- Target deployment of 60 to 150 PCIe400 during LS3 (2026-2028) for LHCb
- Possible interest from Alice, Belle II and CTA collaborations

Project framework

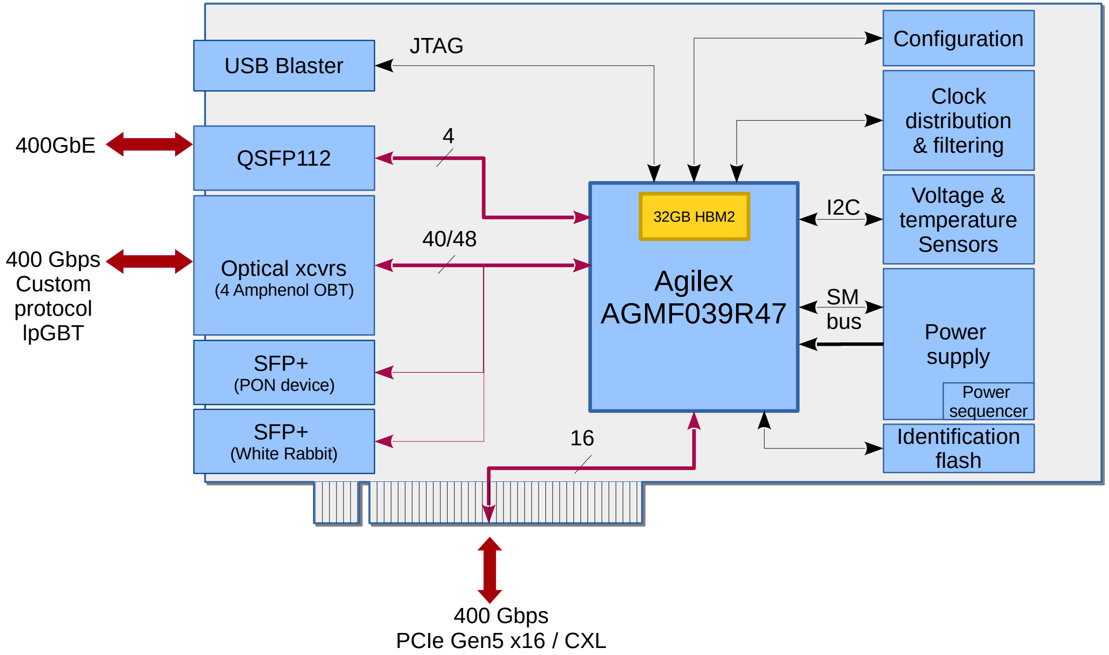
- R&T IN2P3 funded until end of 2024 (Design and prototype phase)
- Involved in 3 ECFA DRD7 projects (7.3b Timing distribution techniques, 7.5b No backend, 7.5c Generic backend)
- Production phase should be pursued through LHCb master project

**under discussion*

PCIe400 overview

Build around the following

- FPGA : Agilex 7 M-series AGMF039R47A1E2V
  - 32GBytes HBM2e 2.6Tbit/s BW
  - ARM Cortex-A53 quad core @1.2GHz
- Optical I/O :
  - up to 48x up-links and down-links at up to 26Gbps,
  - up to 2x SFP+ bidirectional 10Gbps links
  - Optional 1x QSFP112 bidirectional (4x112Gbps link)
- CPU I/O : PCIe Gen 5 x16
- Jitter filter PLL
  - Intrinsic jitter 100fs RMS
  - Versatile clock tree
- Configuration modules and monitoring sensors on-board

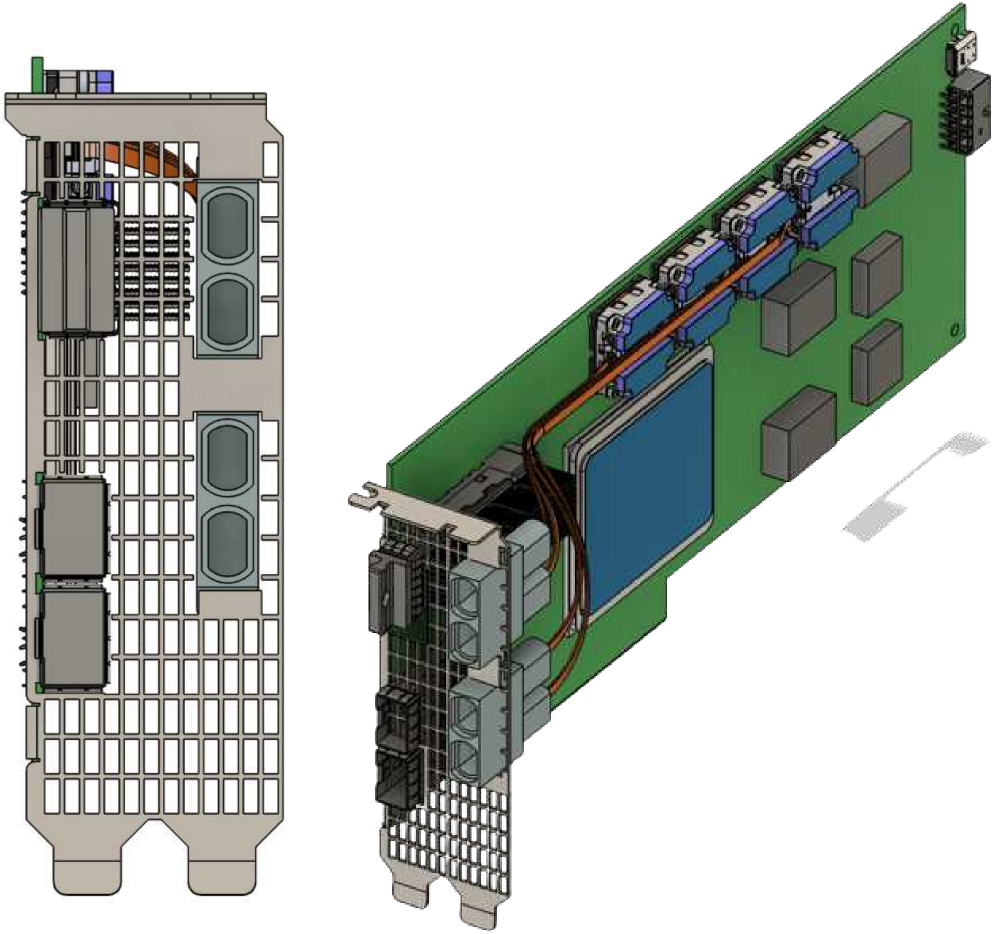


PCIe400 synoptic

Hardware overview

Form factor

- GPU Host server specifically qualify GPU → double width, ~270mm (<full length : 312mm)
- High number of I/O on front plate compared to GPU → high constraints on cooling solution

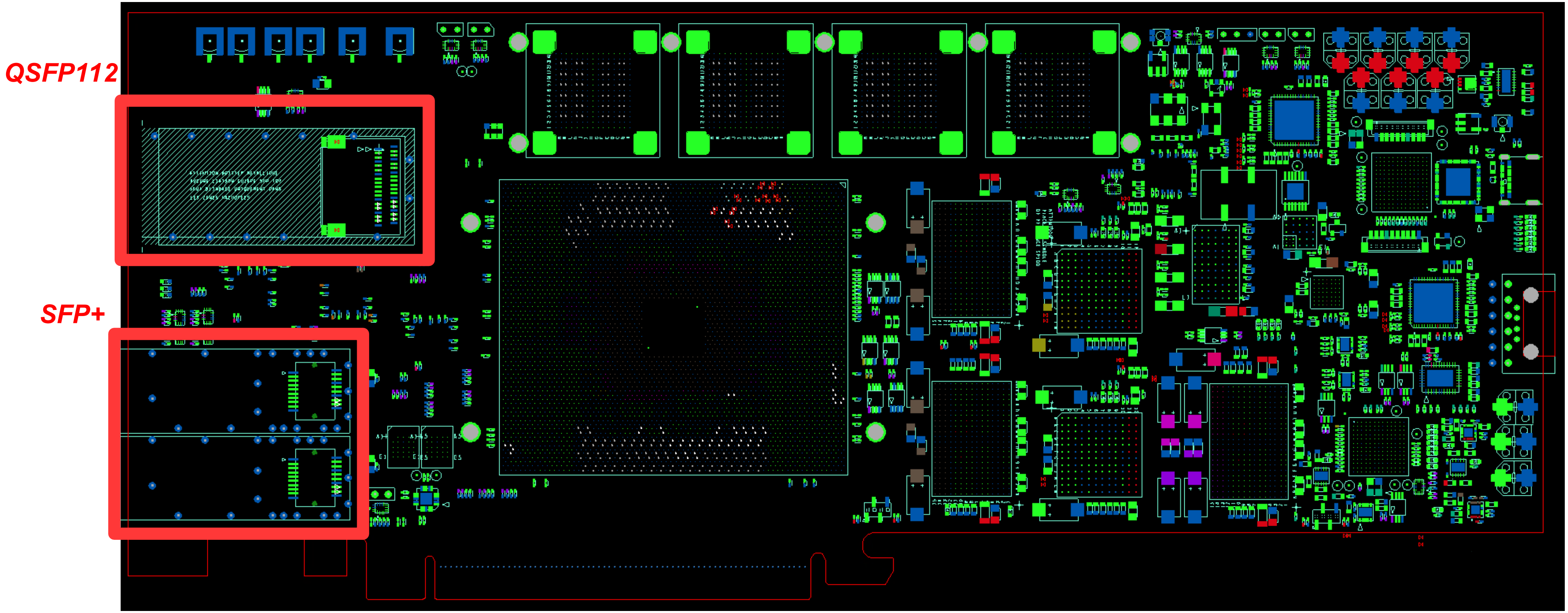


PCIe400 3D model

Hardware overview : Placement

Top

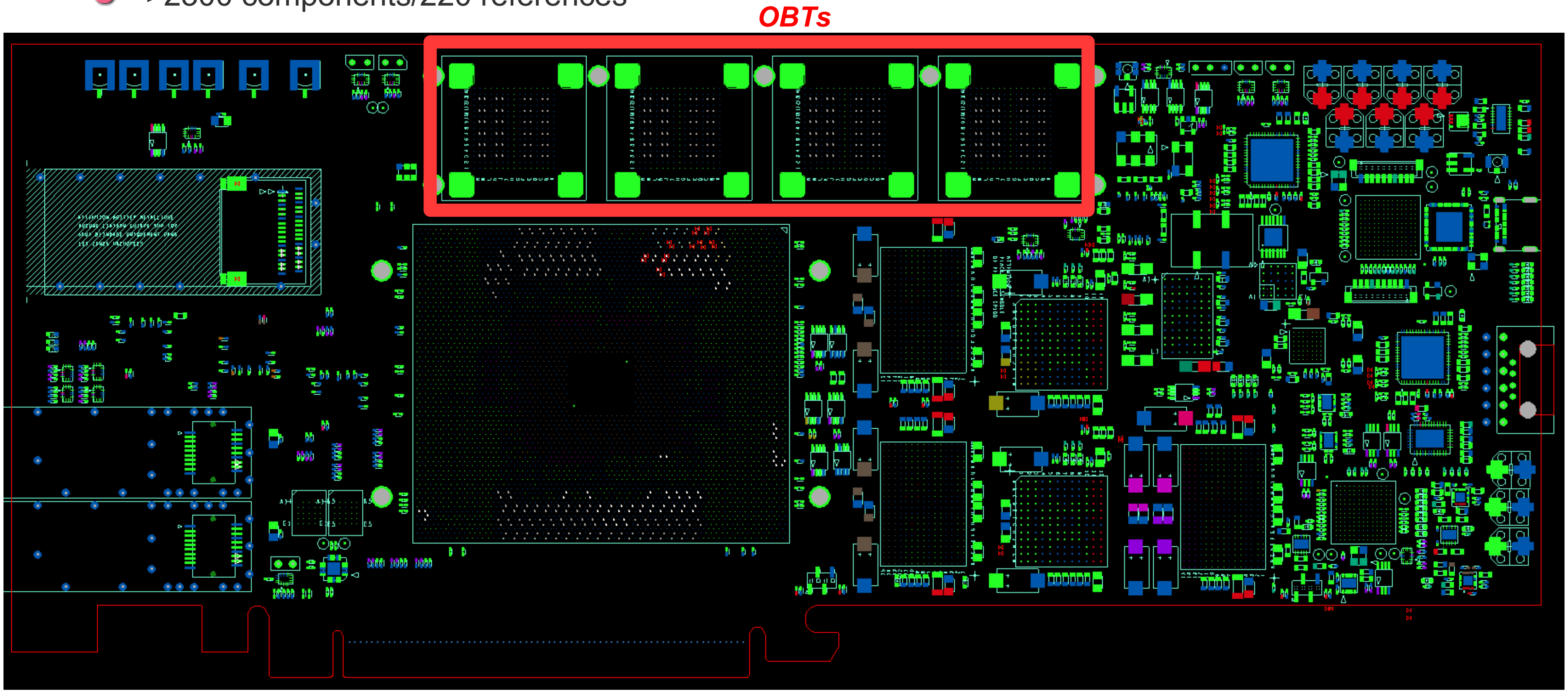
- PCB dimension : 270 x 100mm
- >2300 components/220 references



Hardware overview : Placement

Placement top overview

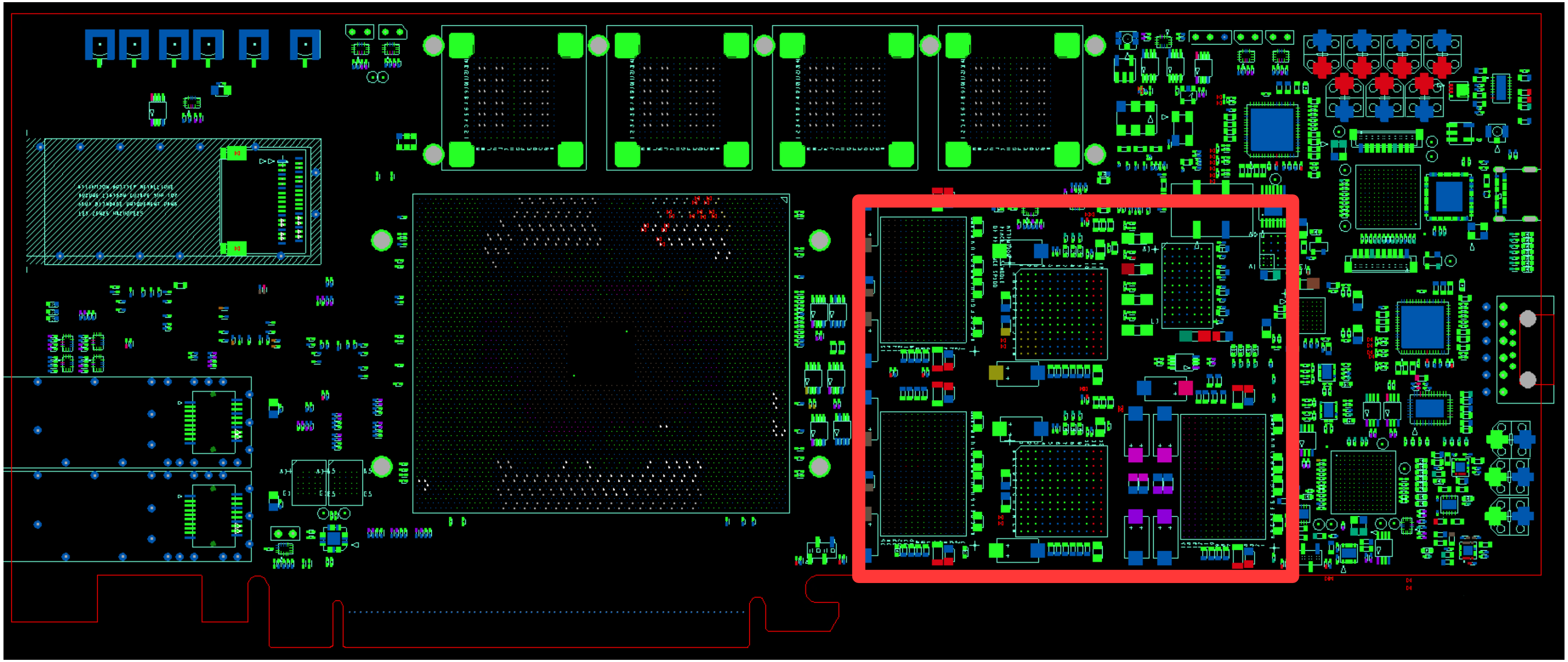
- PCB dimension : 270 x 100mm
- >2300 components/220 references



Hardware overview : Placement

Placement top overview

- PCB dimension : 270 x 100mm
- >2300 components/220 references

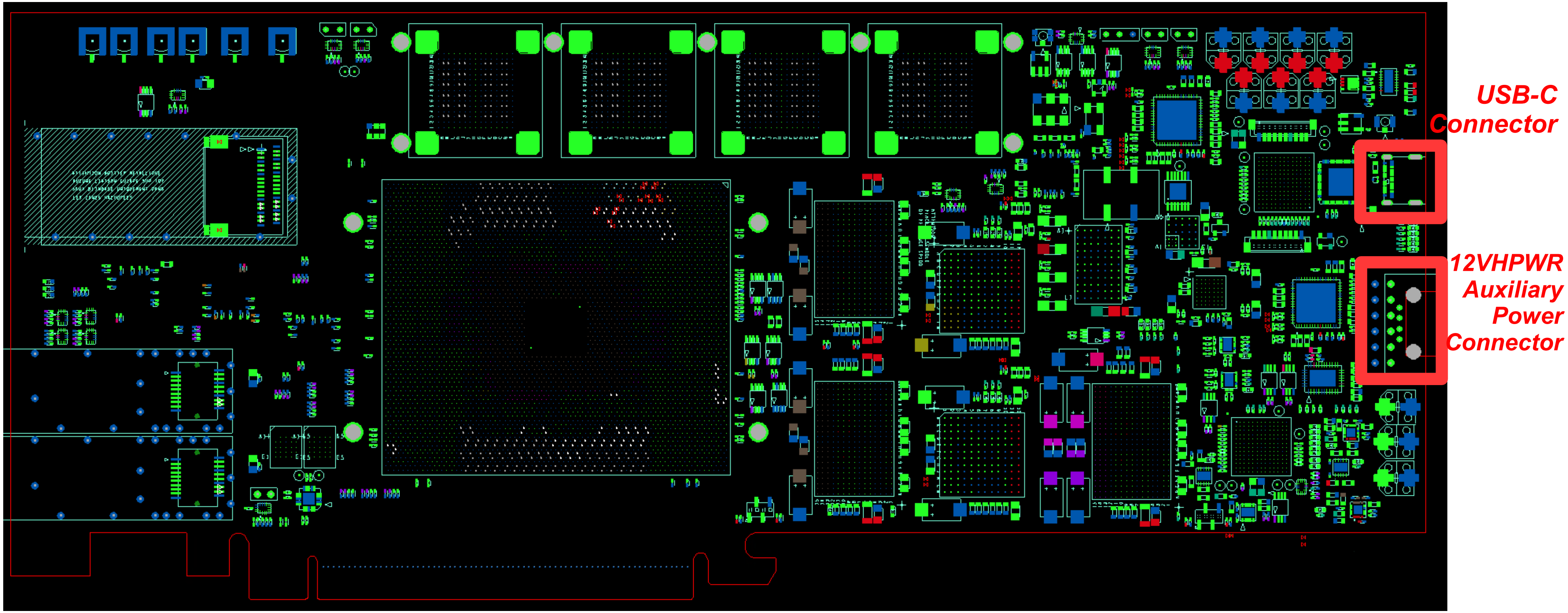


Power Supplies

Hardware overview : Placement

Placement top overview

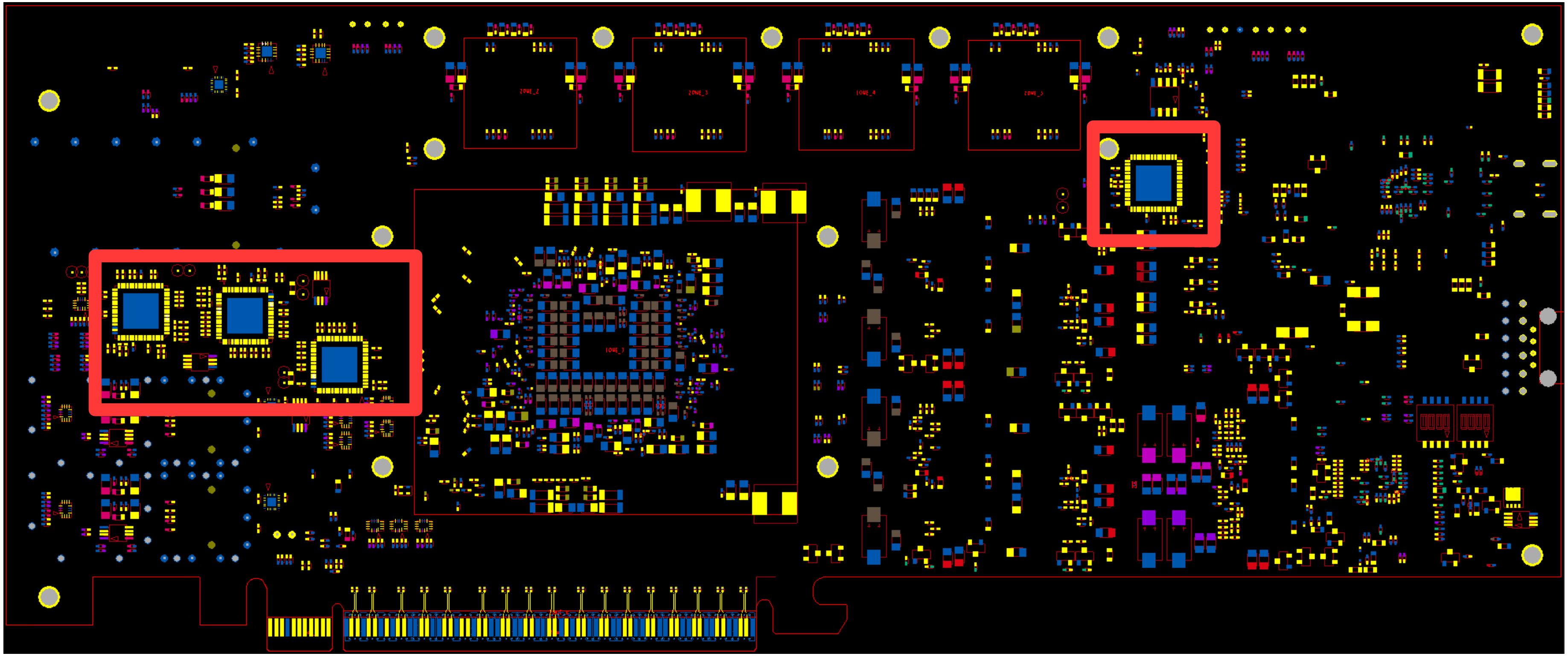
- PCB dimension : 270 x 100mm
- >2300 components/220 references



Hardware overview : Placement

Placement bottom overview

- PLL and components that need to be accessible for debug



White Rabbit jitter cleaner PLL

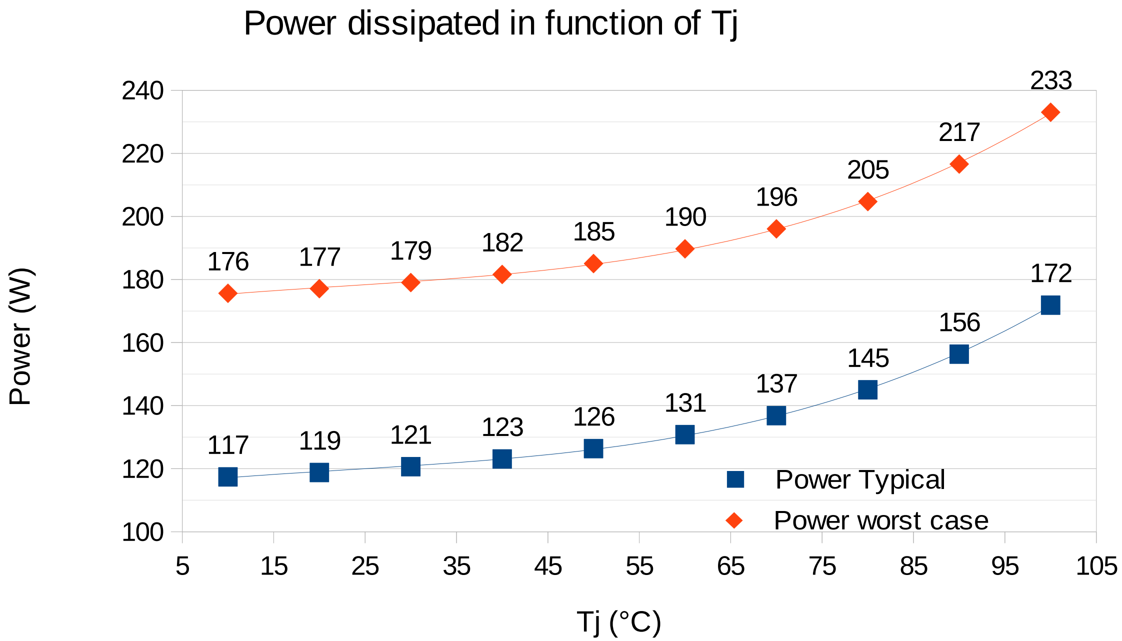
Jitter cleaner PLLs

Technical developments

Power dissipation

FPGA total power dissipated (TDP)

- Estimation at early stage with limited gateware inputs from developers
→ use of statistics from current system
- Estimated between 120W to 220W
- Need for high performance cooling solution



Cooling solution

Air cooling study

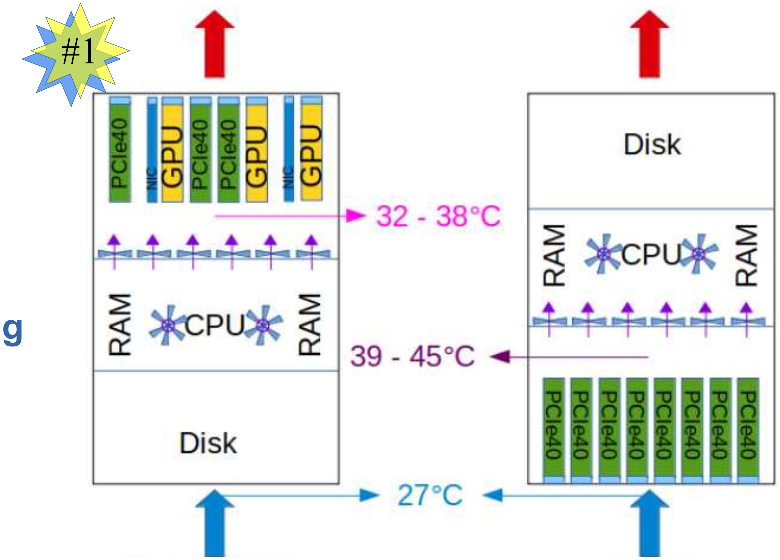
- 2 Airflow architecture identified
- FPGA TDP 160W nominal, 220W worst case
- Optical transceivers must be studied because high power concentration, placement constraints

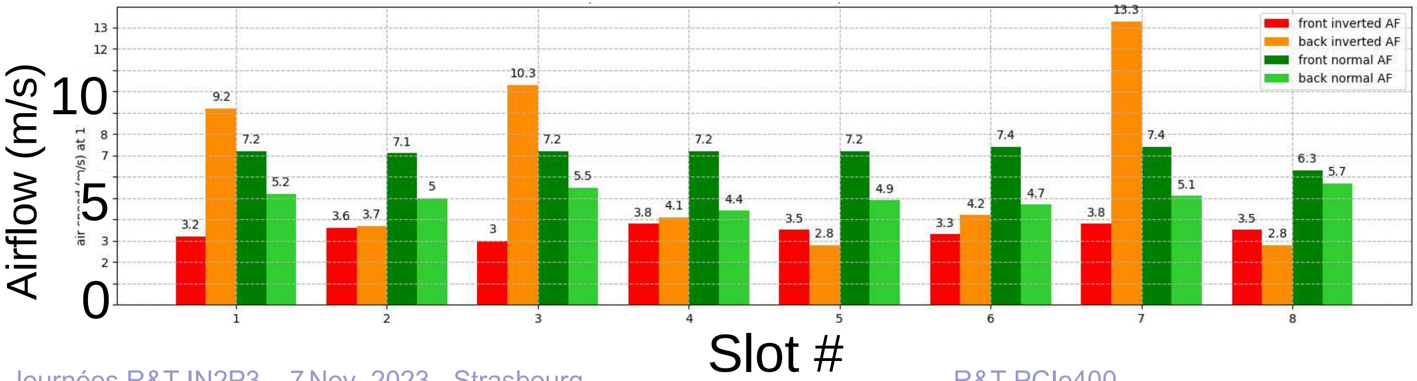
Design of a thermal mock-up instrumented to measure server cooling capacity

- Measure between 4 to 5.5m/s at maximum fan speed

CFD simulation in house

- Model of a vapor chamber using COMSOL (thanks to LAPP)
- Studies to optimize heat-sink (active/passive solution, length)





Retained solution for prototype

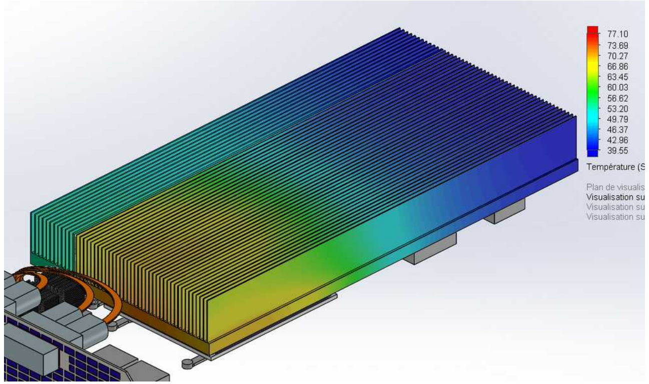
Design and manufacture outsourced with a french company in Grenoble

- Heat-pipe heat-sink with skived fins

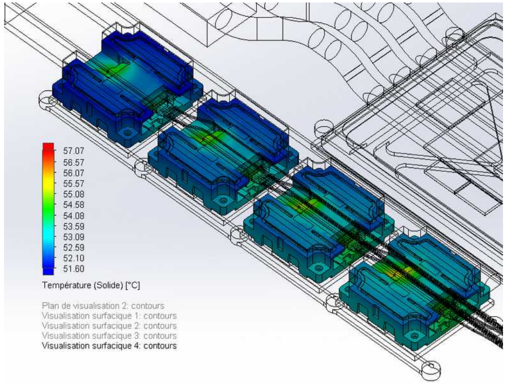
Nominal performance validated in simulation @ 38°C ambient and 5m/s airflow

- FPGA is maintained at 85°C with 160W
- OBT are maintained <60°C at maximum TDP (7.5W)
- QSFP112 is maintained at 75°C with 12W dissipation
- SFP+ are maintained <60°C

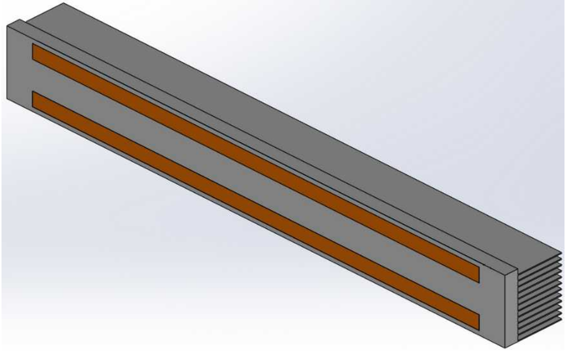
Final cooling solution will be decided after tests on prototype

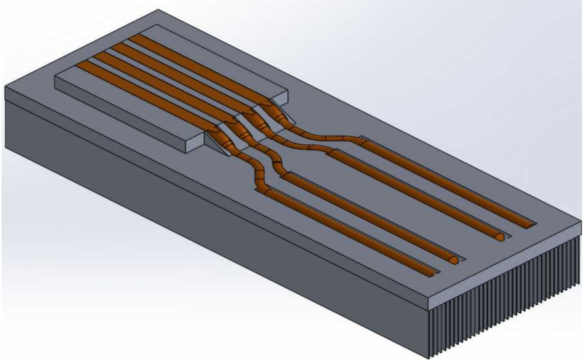


Heat-sink heatmap


57.07
56.57
56.07
55.57
55.08
54.58
54.08
53.59
53.09
52.59
52.10
51.60
Température (Solide) [°C]
Plan de visualisation 2: contours
Visualisation surfacique 1: contours
Visualisation surfacique 2: contours
Visualisation surfacique 3: contours
Visualisation surfacique 4: contours

OBT chassis heatmap close up





Heat-sink design with heatpipe

Power integrity simulation

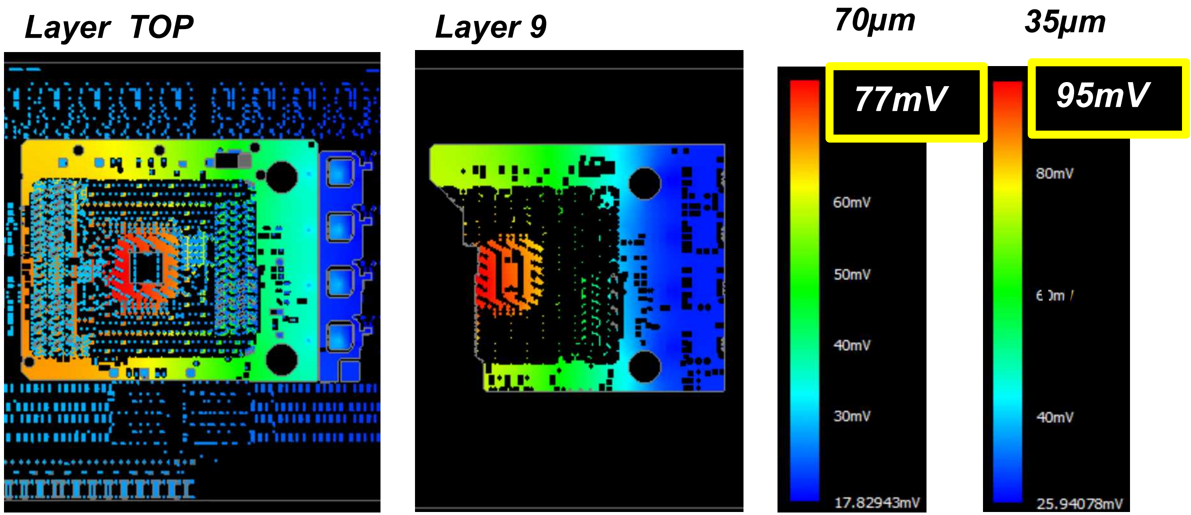
FPGA core power supply

- >100A at 0.8V±1%

Possibility to reduce power planes thickness from 70µm to 35µm ?

- Simulation of FPGA core voltage rail (0.8V @200A) using Intel FPGA Agilex I-series layout (4 power stage + controller)

Voltage drop in power planes



Power dissipation within power plane

Power dissipated	70µm	35µm	Δ
Layer TOP	9.6W	11.0W	+14 %
Layer 9	4.9W	6.8W	+38 %
TOTAL	14.5W	17.8W	+23 %

PCB T° rise	30	17

Simulation with PCB thermal model gives a ~30°C temperature rise in PCB 35µm (T_G = 220°C)

- At early design stage with final planes geometry, 2 internal layers 70µm + 2 external layers ~50µm gives security margins

Differential pair routing

108 differential pairs

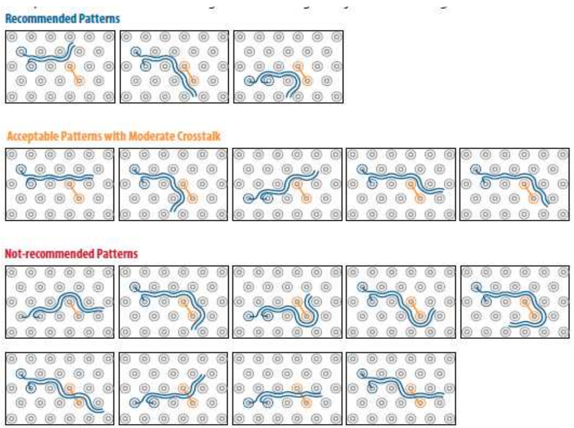
- 112Gbit/s PAM4
- 10, 28 and 32 Gbit/s NRZ

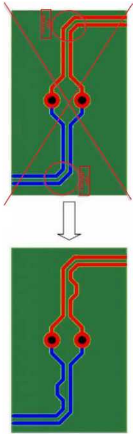
Routing recommendation specified

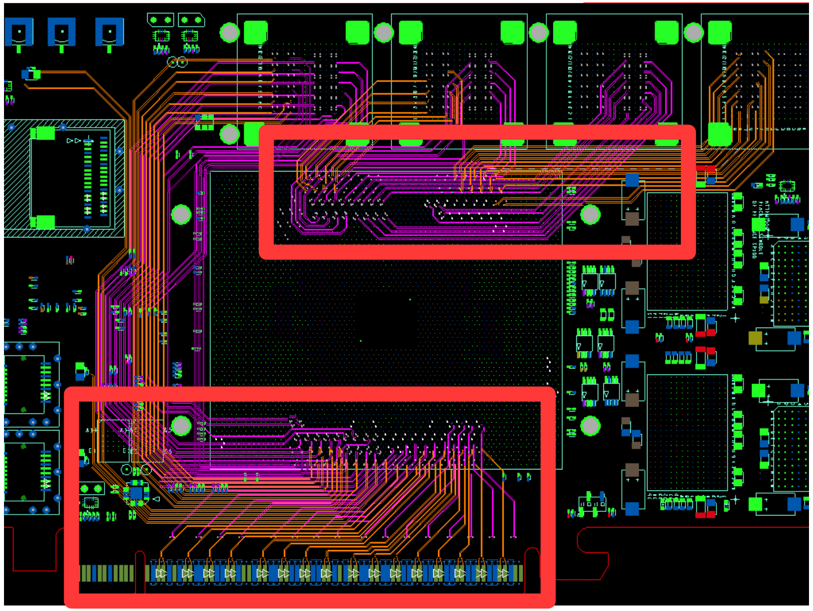
- Controlled impedance ±7% for 112Gbit/s
- Topologies

High density routing due to FPGA form factor and placement constraints

- Near PCIe connector
- In between FPGA and OBT optical transceiver







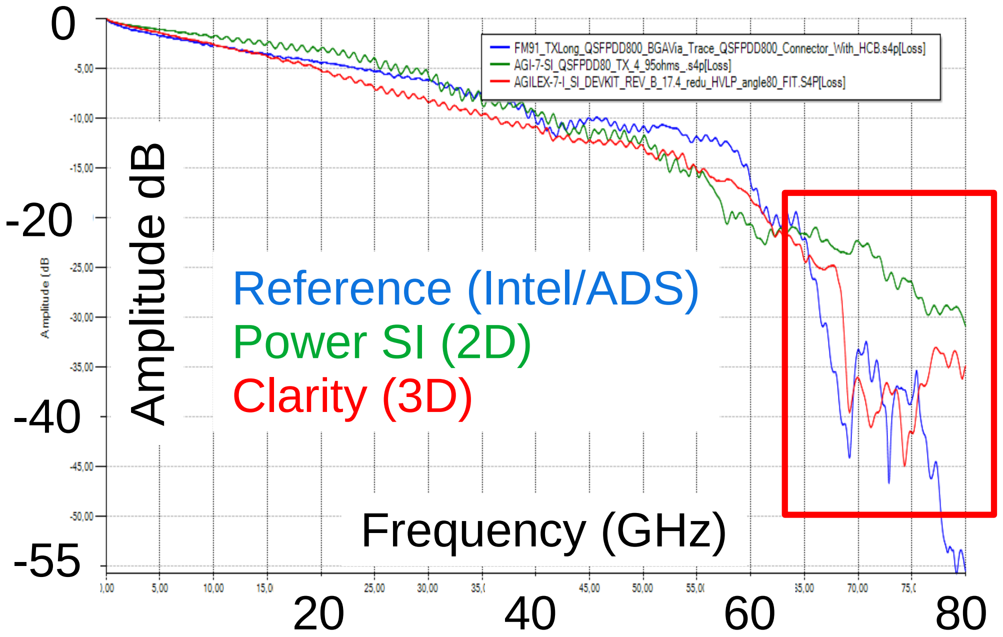
Signal integrity simulation

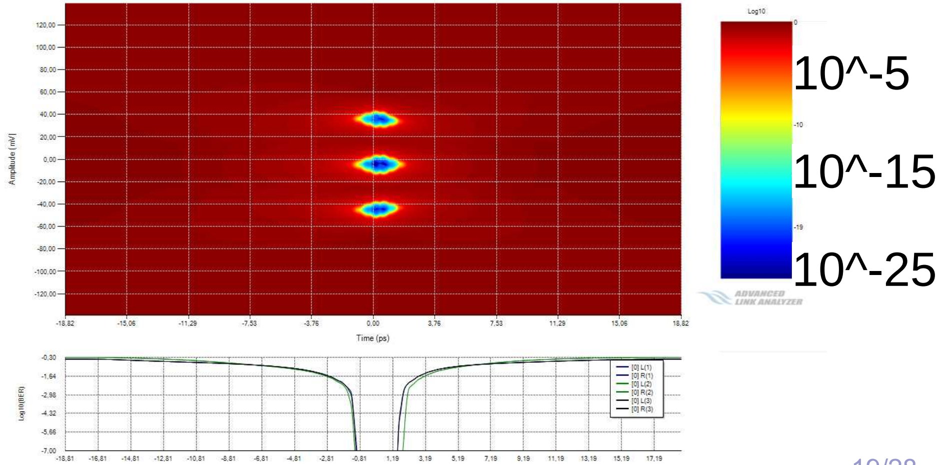
S-parameter extraction

- 112Gbit/s PAM4 → simulation BW 80GHz
- Need for full 3D mesh simulation using Cadence Clarity3D layout → 2D extraction is optimist
- Time consuming simulation but parrallelisable (8h on 64CPU@2.5GHz)

Eye diagram simulation using Intel Advanced Link Analyzer

- The eye only needs to be just opened
- The FEC then insures data valididy





Hardware design : PCB stackup

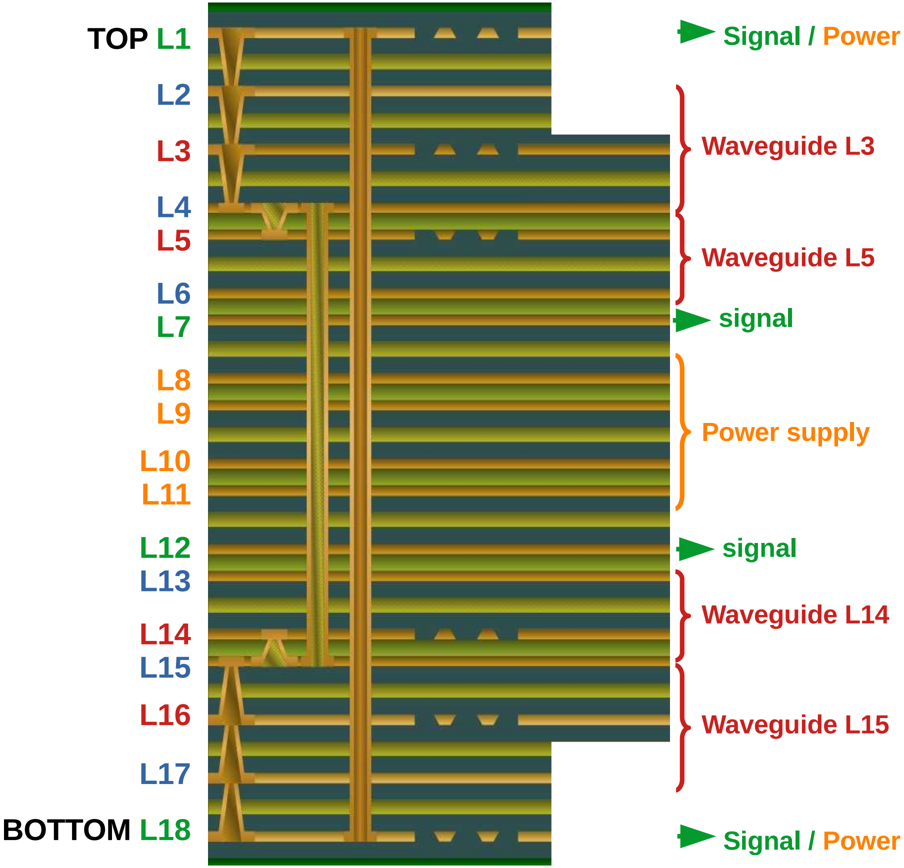
Characteristics

- Dielectric EMC528K (Hallogen free, very low loss)
- HVLP* internal copper foil
- 18 layers, 4 levels of stacked µvias, buried via
- Thickness of 1.45mm at PCIe gold fingers (2 layers removed on each side)
- Total thickness 1.85mm

Manufacturer : Somacis

- Controlled impedance of 7% isn't guaranteed on prototypes, but achievable on production

PCB stackup



Firmware

Golden design

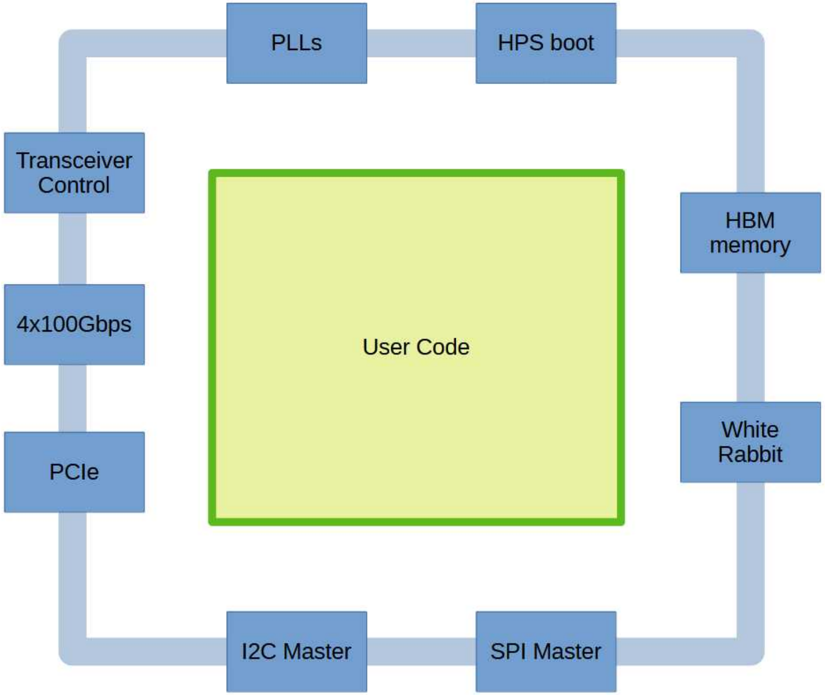
- Validate FPGA pinout with schematics

Low Level Interface (LLI)

- Access board peripherals for configuration, monitoring
- Provide a level abstraction of interfaces for users

Qualification firmware : burning test

- I/O running at full capacity (HBM access, DMA transfer serial links in loopback)
- + FPGA logic filled with random logic pattern



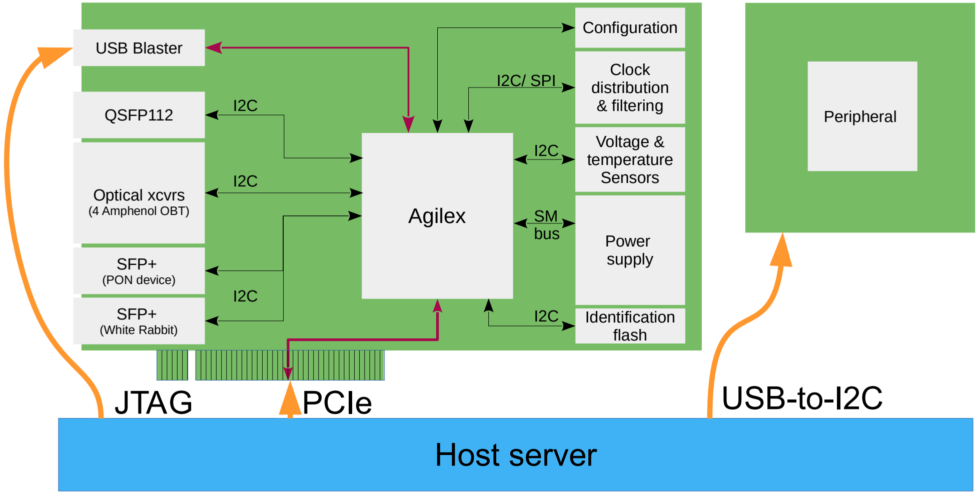
Software

Low level interface

- Provide access to configuration and monitoring registers of FPGA and its peripherals
- Access to board peripherals are centralized on the FPGA
- Several bus available JTAG, PCIe

Software runs on board host server

- Developed with Python and C++ (for interface drivers)
- PyPi package to ease deployment



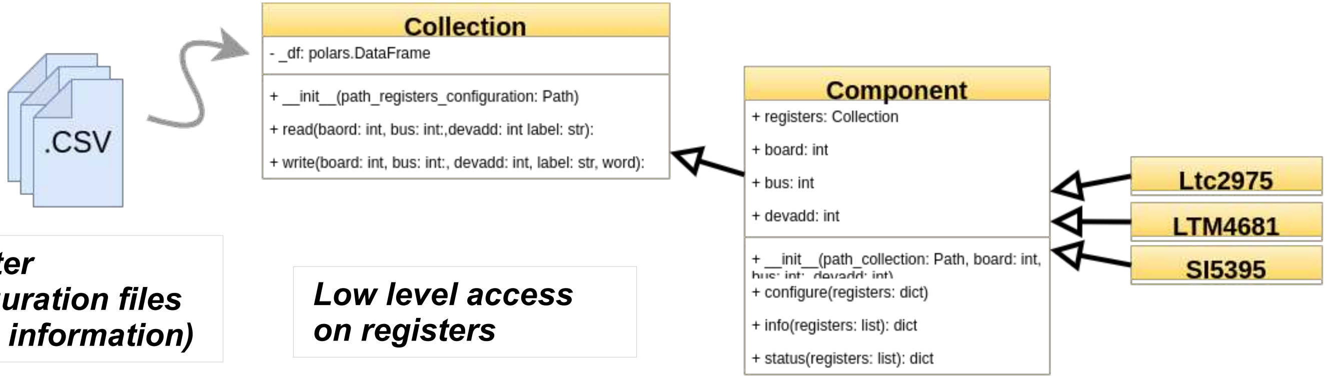
LLI software implementation

Peripheral components

- Each component can be described by a list of registers with limited number of fields

Implementation

- Take advantage of Dataframes and its manipulation methods in order to efficiently access any register fields



Collection

- _df: polars.DataFrame

+ __init__(path_registers_configuration: Path)

+ read(baord: int, bus: int:,devadd: int label: str):

+ write(board: int, bus: int:, devadd: int, label: str, word):

Component

+ registers: Collection

+ board: int

+ bus: int

+ devadd: int

+ __init__(path_collection: Path, board: int, bus: int, devadd: int)

+ configure(registers: dict)

+ info(registers: list): dict

+ status(registers: list): dict

Ltc2975

LTM4681

SI5395

Register configuration files (static information)

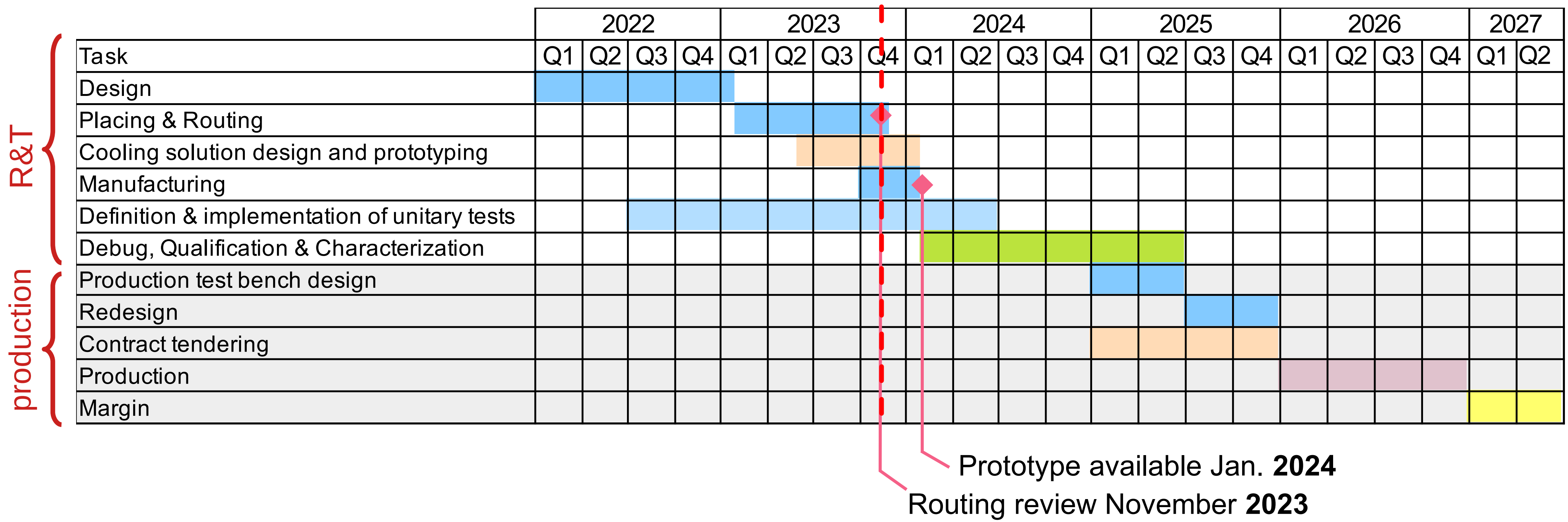
Low level access on registers

Provides common methods to access to batches of registers

Component model : Provide high level specific methods for useful actions

Organization

Planning



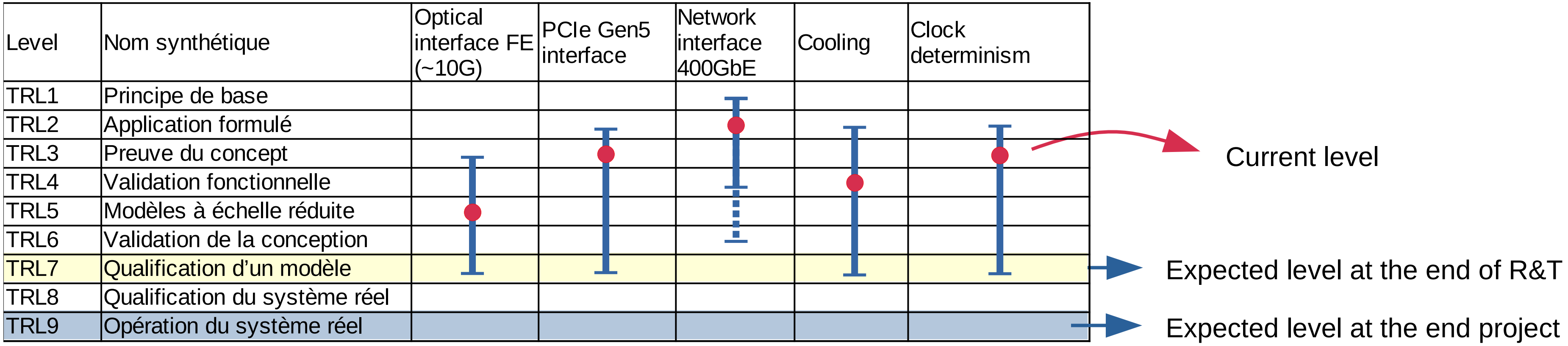
Prototype available Jan. **2024**

Routing review November **2023**

This planning assumes a reinforcement of the team to 4/5 FTE for production phase

- Precise planning of Debug, Qualification and Characterization under construction

Technological Readiness Level Analyzis



Level	Nom synthétique	Optical interface FE (~10G)	PCIe Gen5 interface	Network interface 400GbE	Cooling	Clock determinism
TRL1	Principe de base					
TRL2	Application formulé					
TRL3	Preuve du concept					
TRL4	Validation fonctionnelle					
TRL5	Modèles à échelle réduite					
TRL6	Validation de la conception					
TRL7	Qualification d'un modèle					
TRL8	Qualification du système réel					
TRL9	Opération du système réel					

Current level

Expected level at the end of R&T

Expected level at the end project

Clock determinism (tough spot)

- Specific requirement from HEP application
- FPGA transceivers workaround proposed to / validated by Intel
- A patent might be filed in the future

Network interface (optional)

- Custom implementation of simple network stack (Ethernet, UDP, RoCE*)

*RDMA over Converged Ethernet

Synthesis

PCIe400 is a fundamental development for DAQ system that will pursue beyond R&T framework

- 400Gbit/s output bandwidth per board with up to 48 optical input interfaces
- Baseline solution for LHCb upgrade II, to be published in the Online LS3 enhancement TDR of Q1/2024
- Generic design that can suit several application (Belle II, Alice, CTA)

It also paves ways to explore future DAQ topologies

- 400Gbit/s network interface allowing switch based interconnections or process pipelining between processing unit
- Integration of a white rabbit node for future genearation of precise clock distribution

Participate to the ECFA DRD7 work package

- 7.3c Timing distribution techniques
- 7.5b « No backend », 100GbE at front-end
- 7.5c Generic backend board : benchmark and work toward a common backend design

Many technical challenges yet to overcome

- Requires reinforcement of the team
- An open to newcomers LHCb UII workshop organized to structure the team

Save the date
29 - 30 Nov.
CERN

Merci pour votre attention !